



O'Laughlin & Paris LLP

Public Comment
Bay Delta Plan Workshop 3
Deadline: 10/26/12 by 12 noon

Attorneys at Law

SENT VIA EMAIL

October 26, 2012



Charlie Hoppin, Chairman
State Water Resources Control Board
1001 I Street
Sacramento, CA 95814
commentsletters@waterboards.ca.gov
choppin@waterboards.ca.gov

Re: **Regression Analysis**

Dear Chairman Hoppin:

Since 1988, the California Department of Fish and Game (“DFG”) and other Resource Agencies have put forth the concept that with more flow there will be more returning adult Chinook salmon 2.5 years later. This assertion was initially supported based on a simple linear regression analysis like the one in Figure 5-2 from the 1991 draft State Water Resources Control Board (“State Water Board”) Basin Plan.

This regression analysis was highly disputed in 1991 and in the 1995 Bay-Delta proceedings. The data used was inconsistent, variable and unreliable, particularly since the data collection methodology changed from year to year. Also, a simple linear regression using untransformed, non-normally distributed data is an inappropriate statistical method for elucidating the potential factors influencing adult escapement returns.

Since 1995, the SJTA has consistently made the point to the State Water Board that although adult returns are generally higher 2.5 years after high spring flood flow conditions (>15,000 cfs), the correlation is relatively weak under manageable flow conditions (<8,000). However, you, the State Water Board, have not taken this into consideration during the process that you have undertaken. The draft of the Substitute Environmental Document put maximum flow limits on the Stanislaus, Tuolumne and Merced Rivers. Combined, those limits do not exceed 7,500 cfs, which is within the range of flows where fish returns are highly variable and poorly correlated with flow. The confidence intervals surrounding predicted escapement values based on a logistic regression of the full range of flows over which a much stronger correlation exists, are still too wide to have any useful predictive power.

2617 K Street, Suite 100
Sacramento, California 95816
(916) 993-3962
(916) 993-3688-fax

117 Meyers Street, Suite 110
Chico, California 95928
(530) 899-9755
(530) 899-1367-fax

Mailing Address:
Post Office Box 9259
Chico, California 95927

Therefore, simple linear regression does not provide sound scientific evidence that more flow in the managed flow range will result in more adult returns.

More recently, DFG created and submitted a model (i.e., DFG Salmon Survival Model version 1.6; Marston 2005, revised in Marston and Hubbard 2008) to the State Water Board to further support their assertion that spring flows create more adult returns. In previous comments to the State Water Board, we identified that the “DFG Salmon Survival Model has consistently been found to be inadequate and should not be used” (Demko et al. 2010), and provided supporting documentation (Lorden and Bartroff 2010 Appendix 1 of Demko et al. 2010). Despite receipt of this information, the most recent version the State Water Board’s Technical Report (SWRCB 2012) continues to include DFG’s model.

The State Water Board’s Technical Report (SWRCB 2012) ignores the findings regarding DFG’s model by several independent CALFED peer reviewers, and the State Water Board’s own hearing wherein DFG’s own expert stated, “The model was not ready to be used.” The DFG has continually stated they will update this flawed model according to recommendations by peer review of the DFG’s outflow report that heavily criticized the use of the model as the basis for the conclusions; however, revisions have still not been made after several years.

In this memorandum, we present the results of two different analyses (prepared by Lorden and Bartroff 2012a,b) which confirm our initial assertion that simple linear regression does not provide sound scientific evidence that more flow in the managed flow range will result in more adult returns, and that the DFG Salmon Survival Model is inadequate, not the best available science and should not be used as the basis for consideration of modification of flow regimes in the San Joaquin River Basin. (Lorden and Bartroff 2012a,b) These two analyses examined the following:

1. Simple linear regression analysis of the relationship between San Joaquin River spring flow and fall-run Chinook salmon escapement 2.5 years later (Attachment 1); and
2. DFG’s proposed San Joaquin River Salmon Population model (Attachment 2).

A summary of the key points from each analysis is provided below, followed by the attached report.

Key points regarding the relationship between San Joaquin River spring flow and fall-run Chinook salmon escapement 2.5 years later (Attachment 1).

1. There are a few overly-influential points in the data used to fit the straight-line regression analysis, and conclusions can change drastically from minor changes in the fitting data. Several points represent deviations from the regression that are too large to be consistent with a robust correlation.
2. The data itself shows there is no simple linear relationship between escapement and flow. The most recent 2002-2010 data suggests a negative correlation between these variables, the opposite of the conclusions drawn by the model’s authors.

3. We analyzed the simple linear regression's performance by comparing its predictions of escapement to measured data values and found that the uncertainty in its predictions is so large that the technique has little to no predictive value.
4. The type of regression used is not appropriate for the data, which is indicated through standard statistical measures such as a low R² value and a quantile-quantile plot, indicative of violations of fundamental statistical assumptions.
5. Moreover, we caution against drawing any conclusions of causality between flow and escapement based upon such a regression, which is a well-known fallacy.
6. Environment factors (such as, water temperature, dissolved oxygen and exports), which may aid in understanding escapement, have been ignored in favor of a singular focus on flow. As an example, we have developed a logistic multiple regression using these factors, which describes escapement much better than the proposed model.
7. The simple linear regression does not represent best or widely-accepted statistical practices, and we caution against its use for making any sort of predictions or drawing any inferences about escapement or flow.

Attachment 1: Report on Flow vs. Escapement Model and Environmental Data

Gary Lorden, Ph.D.

Jay Bartroff, Ph.D.

Lordenstats

April 16, 2012

1 The Proposed Simple Regression Model of Escapement on Flow

The proposed simple regression model of SJR escapement on flow has a number of weaknesses. The following four subsections describe weaknesses our analyses have uncovered.

1.1 Evidence Against the Relationship Inferred from the Model Fit

To assess the quality and efficacy of a simple linear regression model of escapement vs. flow, we first performed statistical calculations similar to the ones done by F&G and FISHBIO on the available 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) and 1953-2010 Fall run escapement data. Figures 1-3 show the data, model fit, residuals, and quantile-quantile (Q-Q) plots. Rudimentary straight-line modeling of this kind has been proposed as a useful description of a relationship governing these variables.

If there were such a simple relationship between these variables, that relationship should appear consistently when one partitions the 57 data points into subsets. We have examined two natural ways of doing this, breaking up the data into groups according to time periods and according to magnitudes of flow. In both cases, the results were inconsistent, calling into question the validity of the proposed simple relationship.

Figure 4 shows the same data and straight-line fit in black as in the first plot, Figure 1, but here the 2000-2008 Vernalis Spring flow vs. 2002-2010 Fall run escapement data is shown in red, along with a red line fitted to those data points by the same linear regression method. The 2002-2010 escapement data actually has a *negative* correlation between escapement and flow, and hence the red line has a negative slope. Since these data for the last 9 years constitute the most recent data, it would seem that they provide an important check on the potential value of the proposed linear model in predicting a relationship between flow and escapement in future years. It has been brought to our attention that this period of 2002-2010 for escapement data (and corresponding period of 2000-2008 for flow data) is in fact one in which a new program of water resource management has been in effect.

Figure 5 shows the data and fitted lines when the flow range is broken into 1, 2, 3, or 4 bins of equal sizes. The fitted lines (and hence, the correlation estimates) vary from bin to bin, indicating

that there is not a linear relationship that holds over the entire range of flow values. Note that one of the fits in the fourth row even has negative slope. All of these fits suffer from low R^2 values: The ten plots in Figure 5 have R^2 values in the range [.0043, .41].

Additional doubts about the validity and value of the linear-fit model arose when we noticed that a small number of data points overly influence and inflate the linear relationship between escapement and flow. It is well known that simple linear regression is highly non-robust and can easily be “fooled” by a small number of data points. (See also the discussion of outliers in Section 1.2). When a small number of data points are overly influential, one would expect to see inconsistencies between linear fits made using random subsets of the data points, since the highly influential points will affect some fits and not others. This behavior is observed in Figure 6, where the data were divided into four subsets of equal size at random; each row represents an independent realization of this process. Note that the model fits vary widely and a negative correlation is even found in one subset in the first and second realizations.

1.2 Violations of Model Assumptions

Returning to Figures 1-3, there are several fundamental assumptions of the regression model that seem to be violated by the data.

The model assumes that the observations of the y variable, here escapement, is normally distributed. When this holds, the shape of points in the scatterplot is roughly “football” shaped along the fit line, which is not evident in Figure 1. Another standard way to assess this normality assumption is to examine the Q-Q plot of the residuals, which compares their distribution to the assumed normal distribution. If normality holds, then the points in the Q-Q plot should lie close to the dotted line in the third plot. The fact that they are not close in Figure 3 is evidence of non-normality.

Another assumption of the model is that observations of the y variable are subject to random variations whose scale is constant and which average out to zero. When this holds, the residual plot should appear as roughly a uniform cloud of points, symmetric around the horizontal dotted line. That is not the case in Figure 2, which on the contrary indicates both a bias (non-zero average) and a non-constant scale of variations. Moreover, the numbered points in Figures 2 and 3 are outliers – points that represent deviations from the linear model that are too large to be consistent with that model.

Finally, we note that the model fit in Figure 1 has an R^2 value of .27. R^2 , the coefficient of determination, is the square of the correlation coefficient and thus takes values between 0 and 1. R^2 is a measure of goodness of fit of the model and, more specifically, is the fraction of the variation in the y data that is considered to be “explained by the linear fit” on the x data. A value of .27 is generally considered quite low and indicates that this proposed model does not capture a meaningful relationship between the two variables.

1.3 Lack of Predictive Power

As would be expected from its poor fit of the available data, particularly in the most recent time period, the linear model seems to have very little predictive power. A standard way to assess the usefulness of a fitted model is to calculate and examine so-called “prediction intervals” computed from it. These are confidence intervals for future observations, calculated so that they should be correct at some prescribed confidence level. Table 1 contains prediction intervals calculated at the

95% confidence level from the linear model fit to the 1951-1999 Vernalis Spring flow (Mar. 15 - June 15) and the 1953-2001 Fall run escapement data, and compared with the actual 2000-2008 flow and 2002-2010 escapement data; in the table and Figure 7, “year” refers to escapement year. These prediction intervals are extremely wide – too wide to have any useful predictive power. For example, with the exception of one year, the upper prediction for each year is larger than any escapement measurement made in the entire combined data sets. (The largest escapement measurement was 80,000 while the 2004 upper prediction was 79,324). In spite of their extreme width, the prediction intervals for two years – 2007 and 2008 – do not in fact contain the actual escapements observed in those years. Figure 7 contains a graphical representation of the prediction intervals and the actual 2002-2010 observations.

Table 1: Predictions and 95% confidence prediction intervals. Values in bold are violated by observed escapements.

Year	Vernalis flow (avg. over daily values 2.5 yrs prior)	Escapement	Predicted Escapement	Lower Prediction	Upper Prediction
2002	4811	25666	12266	1288	116836
2003	3185	11144	9197	967	87447
2004	2611	10319	8006	841	76194
2005	2707	6376	8211	863	78128
2006	2476	4169	7715	810	73458
2007	10597	1241	21287	2196	206354
2008	25545	2229	39339	3896	397268
2009	2829	1323	8468	890	80544
2010	2376	2423	7498	787	71416

1.4 Inferential Problems

Because linear regression analysis is so widely used, a number of mistakes and fallacies that occur frequently in their interpretation are well known. Two that are relevant here are the Ecological Fallacy and the Correlation/Causation Fallacy.

The Ecological Fallacy refers to making inferences at the individual level based on regression analysis performed at a subgroup level. This typically occurs when data are averaged or combined over a subgroup before fitting a regression model. This can lead to fallacious conclusions because averaging reduces variation and therefore can falsely inflate the strength of linear relationship, or make one appear when in fact there is a more complex relationship– or no relationship at all. The current proposed model is in danger of this because the flow data are averaged over two months before performing the regression fit, a very crude form of data reduction in this setting that suppresses a large source of natural variability. The proposers of this model have the responsibility to show that the variation lost in averaging does not affect the inferred relationship.

Another relevant fallacy is the Correlation/Causation Fallacy, in which an estimated correlation in a regression analysis is mistaken for causation– i.e. that the variables have a genuine cause-and-effect relationship. Although a robust model fit can indicate a possibility of causation, that is not the case for the sort of linear model proposed between flow and escapement, which is highly non-

robust in light of the inconsistencies cited in Section 1.1 and the violations of model assumptions cited in Section 1.2. The proposers have not shown that the estimated correlation corresponds with a causal relationship.

2 Environmental data

Figures 8-13 contain boxplots of the available environmental data, before any averaging occurs. Figure 14 contains scatterplots of this data, on the log scale, after being averaged over the period Mar. 15 - June 15 for each year of available data. In these scatterplots the escapement data were paired with the corresponding variable from two years prior. The temperature data in Figure 14 is hourly, back to 1999, and was obtained from the California Department of Water Resources webpage.

Other than Vernalis flow, there is an overall scarcity of environmental data available, and what exists is further compressed by the yearly averaging. We suspect that this is one reason for the focus on Vernalis flow by F&G as an “explanatory” variable for escapement. For example, it is clear that water temperature may have a large affect on escapement. However, hourly water temperature data is available only back to 1999. After averaging and matching up with escapement data two years later, this results in only nine data points corresponding with the 2001-2009 escapement data. This is a small amount of data to develop any sort of meaningful model. Note also that even if other environmental variables had more data available, any model that includes temperature would be restricted to using only these nine years.

We have fit a multiple regression model of $y =$ SJR escapement (on the logarithmic scale) on the variables

$$\begin{aligned}x_1 &= \text{Vernalis temperature} \\x_2 &= \text{Mossdale dissolved oxygen} \\x_3 &= \text{Mossdale temperature} \\x_4 &= \text{CVP exports} \\x_5 &= \text{SWP exports},\end{aligned}$$

all on the logarithmic scale, depicted in Figure 14. Quadratic terms were included for the Vernalis and Mossdale temperature variables since it is expected that extreme temperatures, both low and high, tend to reduce escapement. The least squares model fit is given by

$$y = -14092.5 + 777.7x_1 - 113.0x_1^2 + 14.2x_2 + 5909.3x_3 - 681.9x_3^2 - 4.2x_4 + 4.6x_5,$$

and has an R^2 value of .6. Though the small number of data points likely causes this R^2 value to be somewhat inflated, this result suggests that one might be able to model escapement in a statistically useful way using multiple variables in addition to flow.

1953–2010 Escapement vs. 1951–2008 Flow ($R^2 = 0.27$)

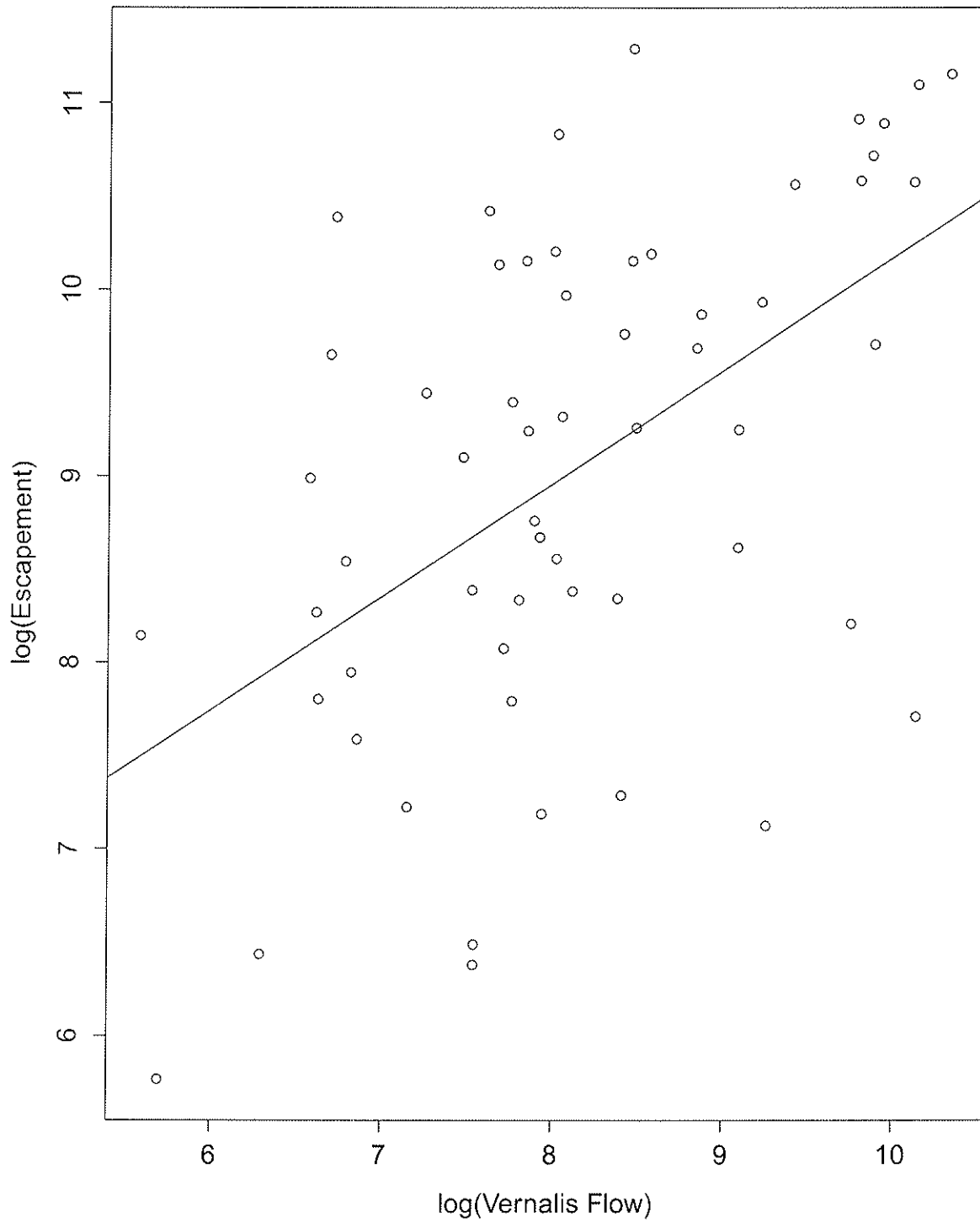


Figure 1: Data and linear model fit for 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) vs. 1953-2010 Fall run escapement data, both on the logarithmic scale.

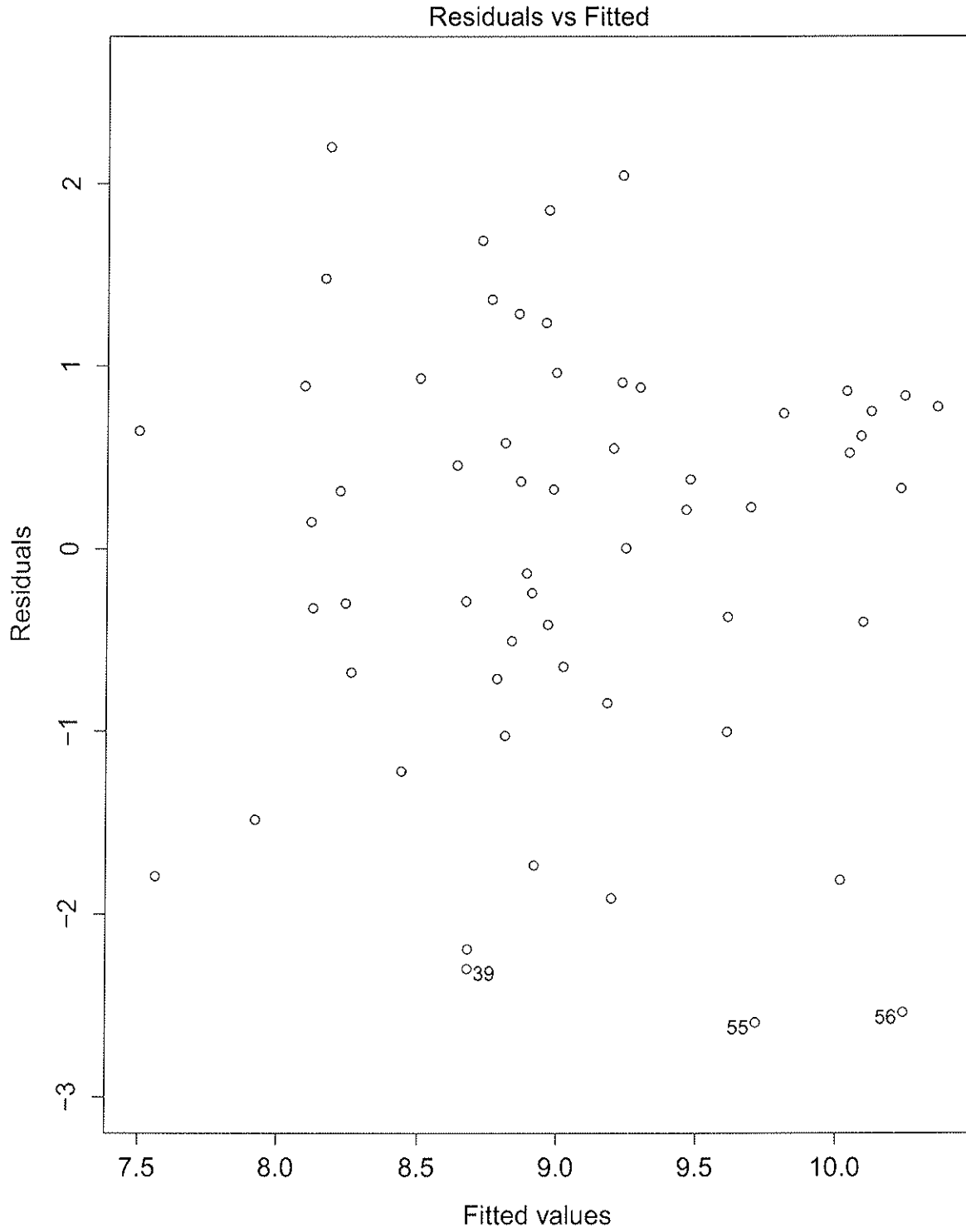


Figure 2: Residuals for 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) vs. 1953-2010 Fall run escapement data, both on the logarithmic scale.

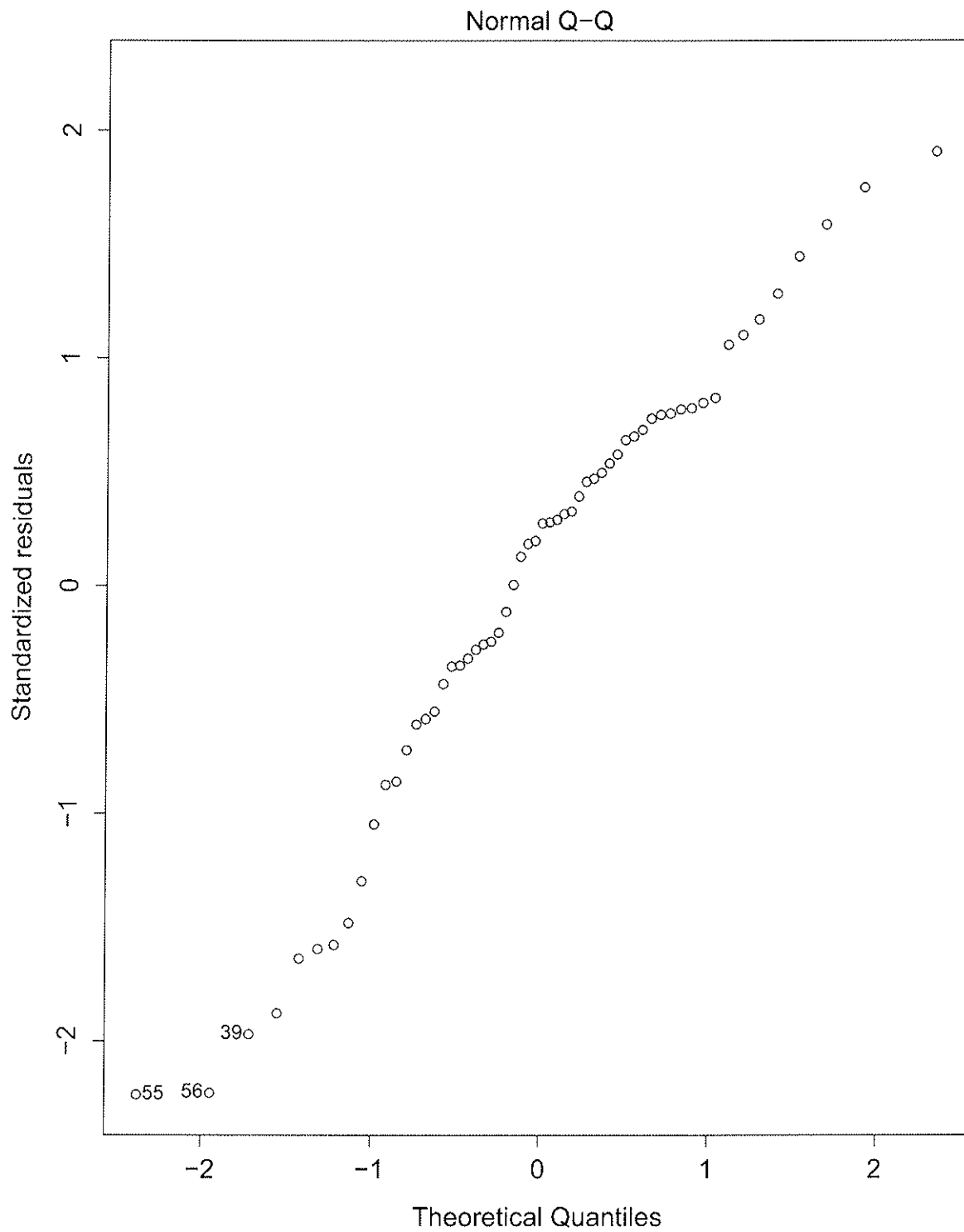


Figure 3: Quantile-quantile plot for 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) vs. 1953-2010 Fall run escapement data, both on the logarithmic scale.

Escapement vs. Vernalis Flow

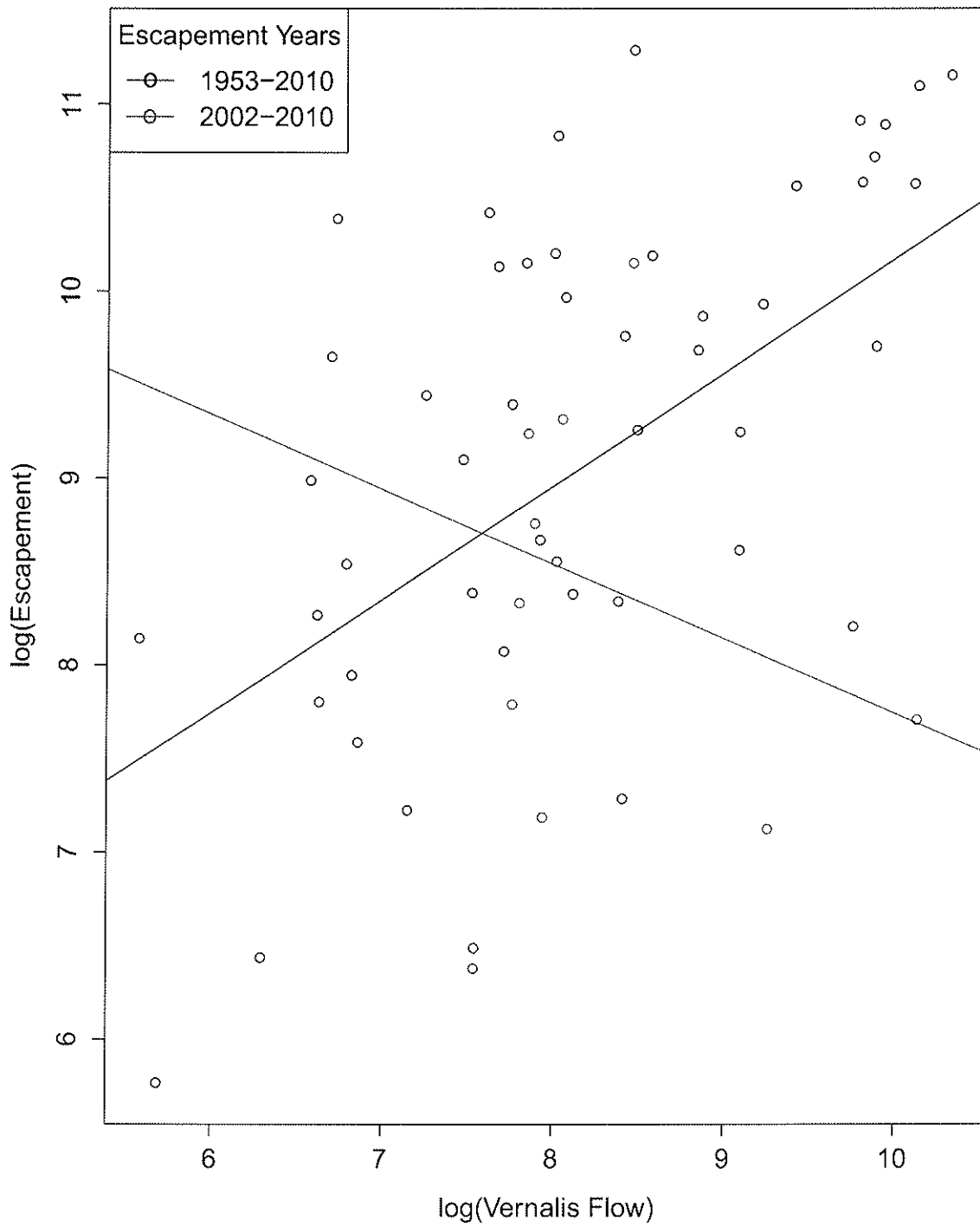


Figure 4: Data and linear model fits for the 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) vs. 1953-2010 Fall run escapement data, and 2000-2008 Vernalis Spring flow vs. 2002-2010 Fall run escapement data.

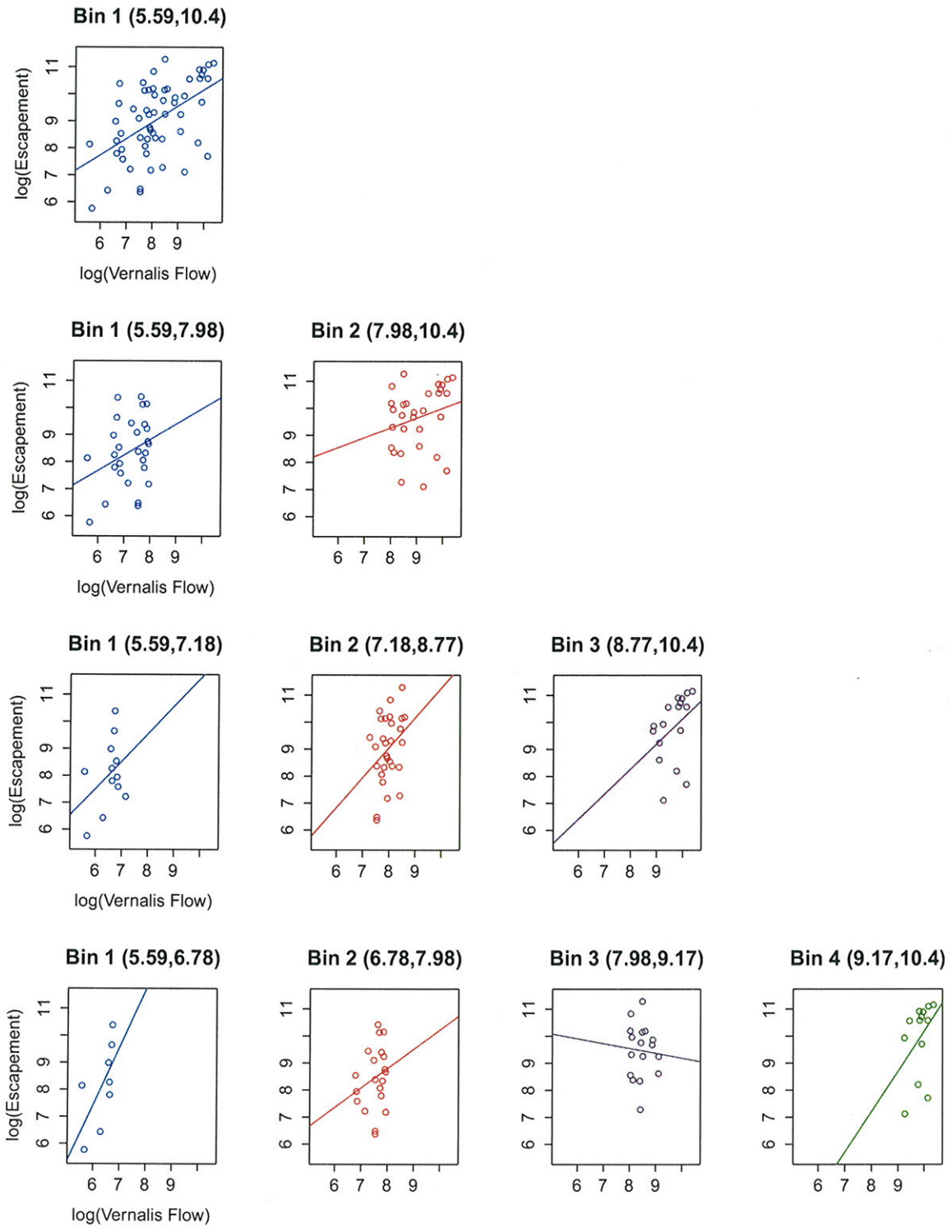


Figure 5: Data and linear model fits for the 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) vs. 1953-2010 Fall run escapement data when flow range is divided into 1-4 equally sized bins (rows 1-4). The R^2 values for these ten fits are all low, in the range [.0043, .41].

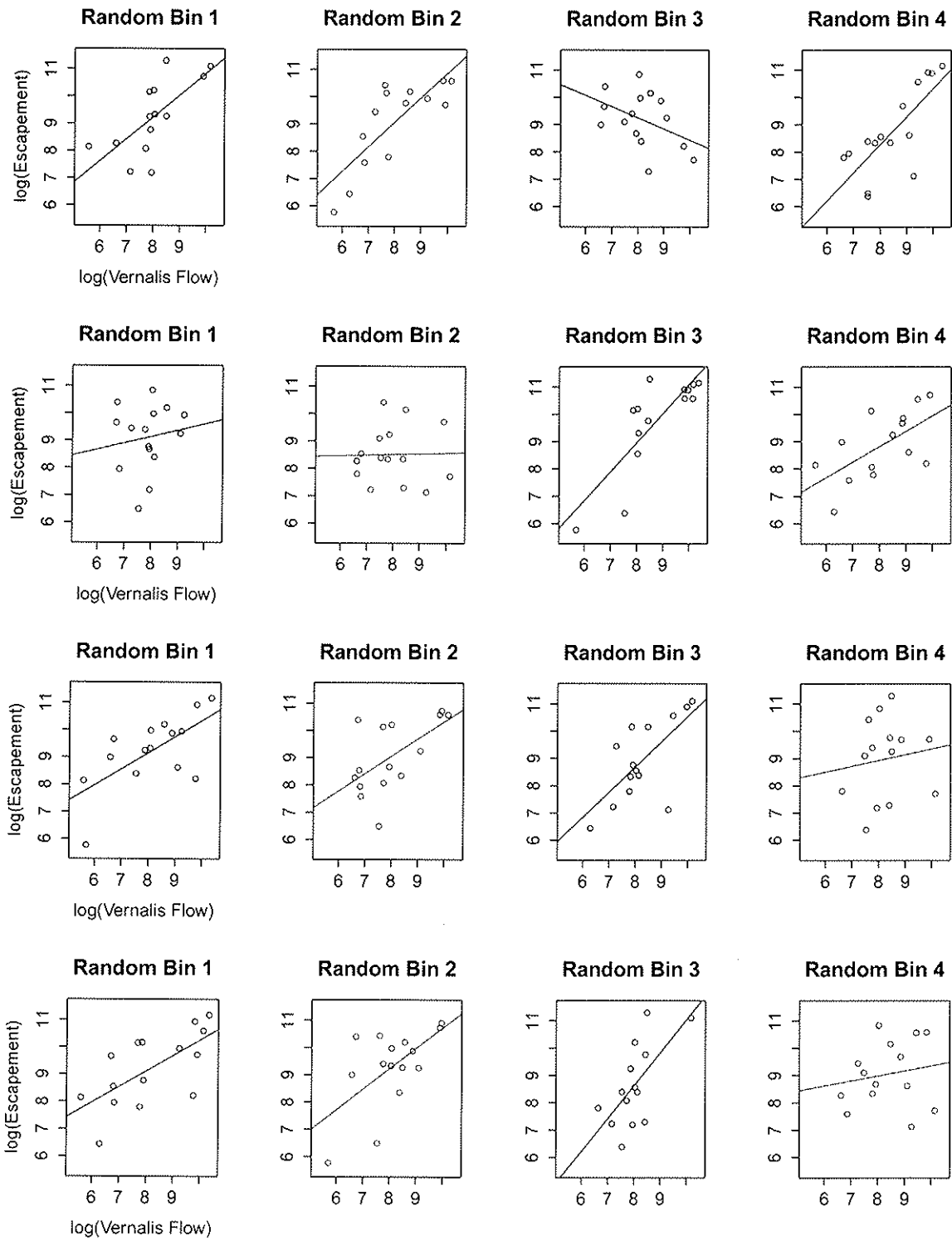


Figure 6: Data and linear model fits for the 1951-2008 Vernalis Spring flow (Mar. 15 - June 15) vs. 1953-2010 Fall run escapement data divided into four subsets at random. Each row is an independent realization.

Pre-2002 95% Prediction Intervals and Post-2002 Obs.

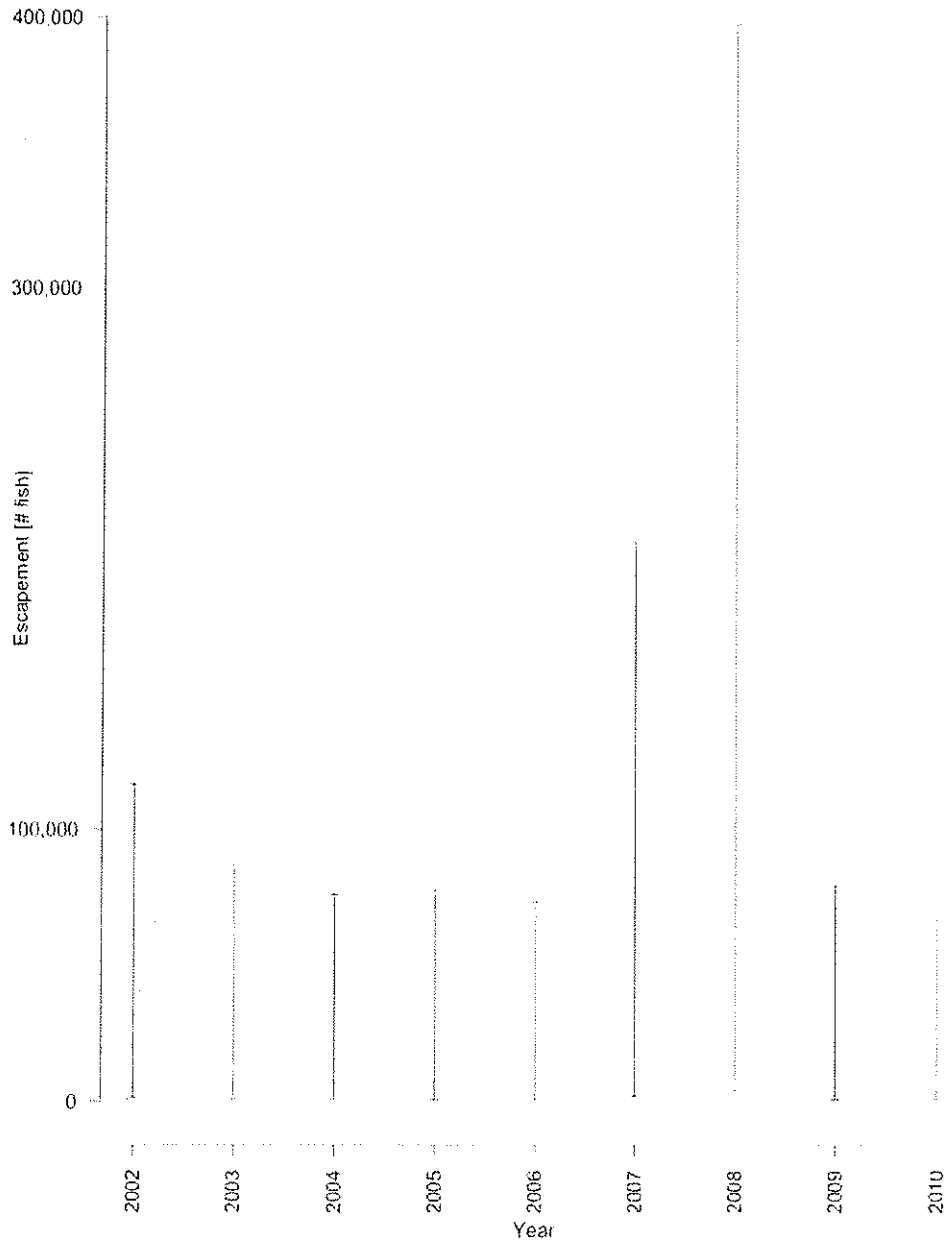


Figure 7: 95% confidence prediction intervals from the 1951-1999 Vernalis Spring flow (Mar. 15 - June 15) and the 1953-2001 Fall run escapement data, and compared with the actual 2000-2008 flow and 2002-2010 escapement data.

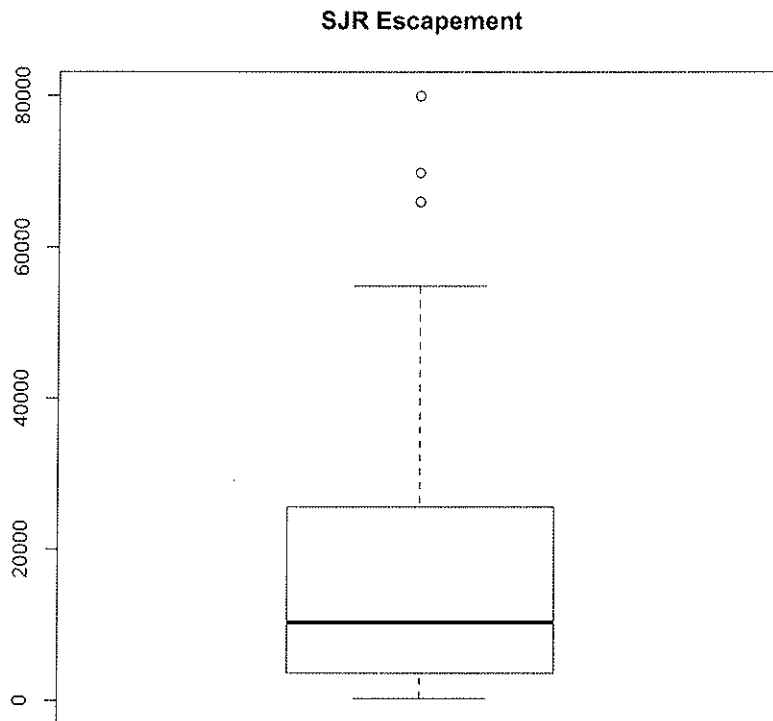


Figure 8: Boxplot of SJR escapement data, 1952-2010.

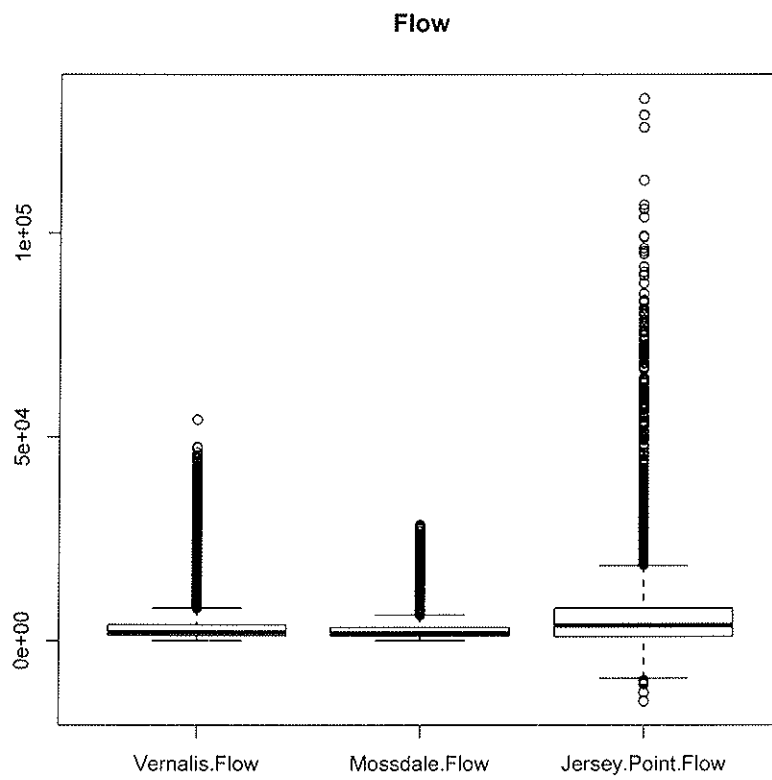


Figure 9: Boxplots of daily flow data.

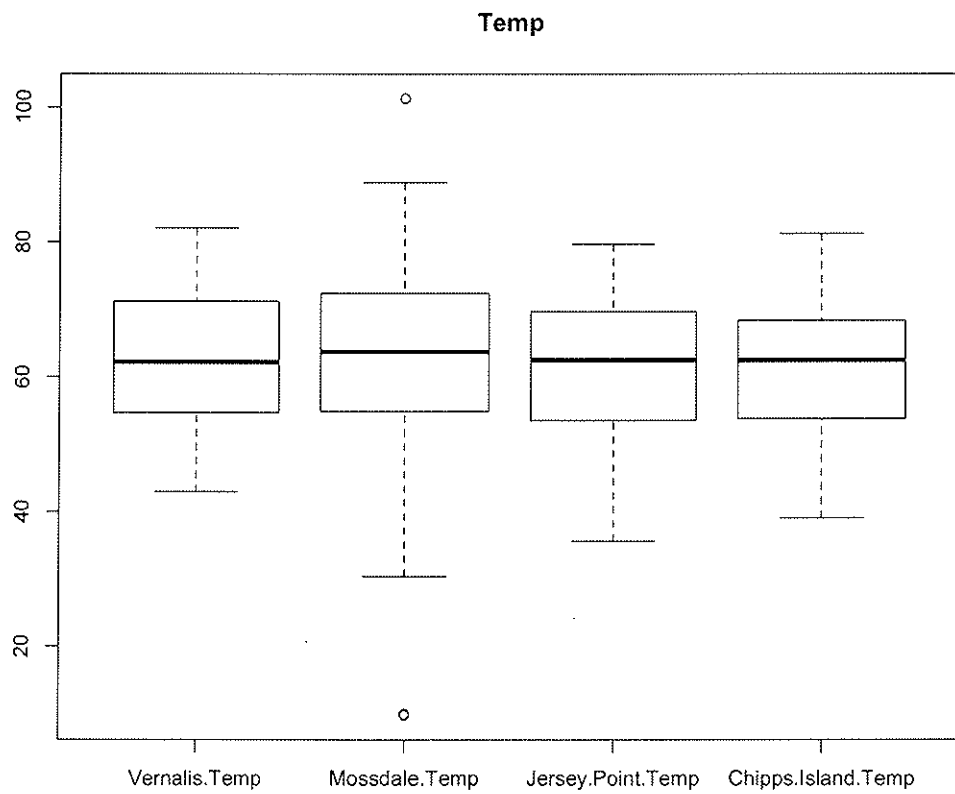


Figure 10: Boxplot of daily temperature data.

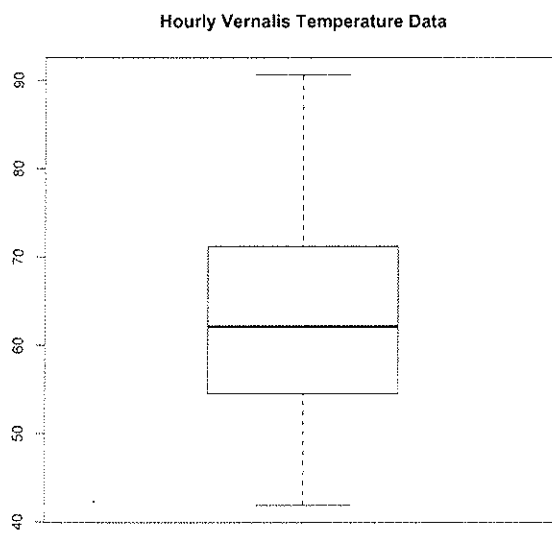


Figure 11: Boxplot of hourly Vernalis temperature data.

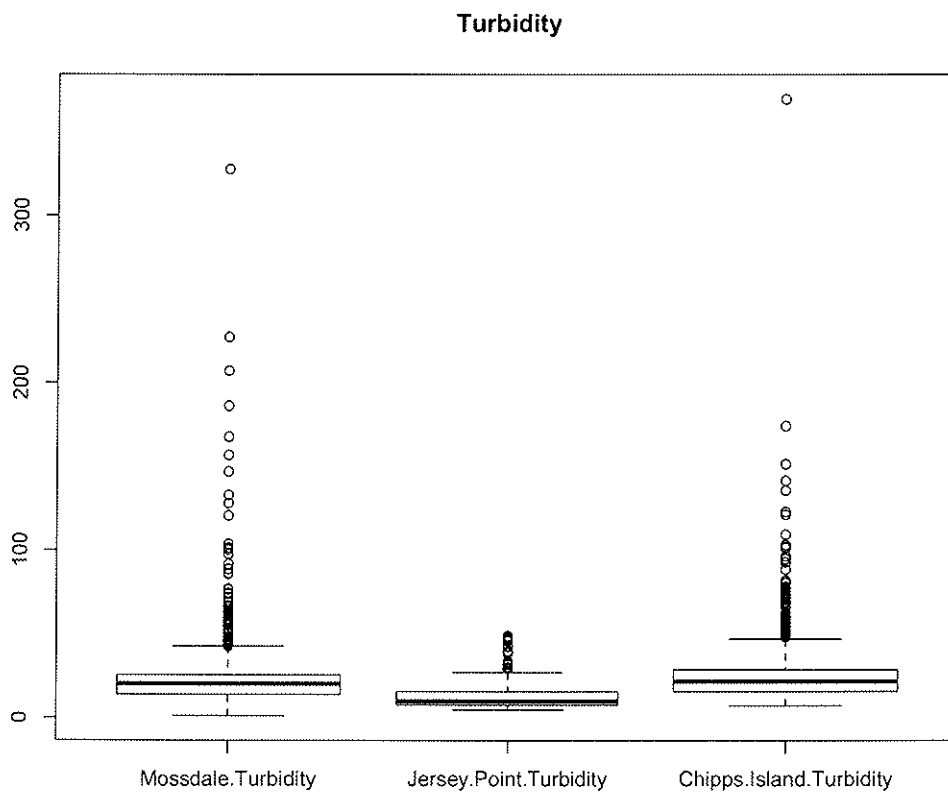


Figure 12: Boxplot of turbidity data.

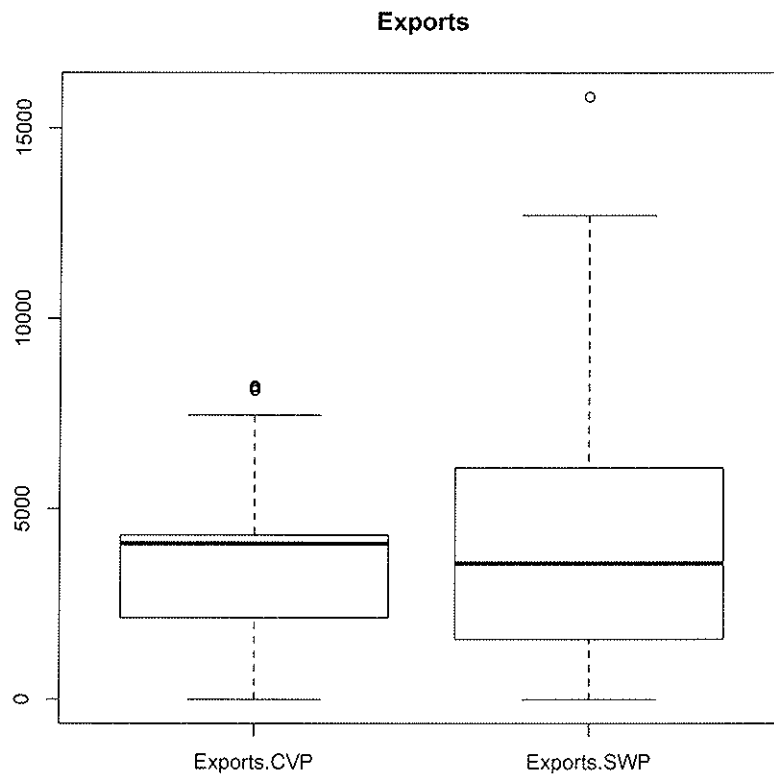


Figure 13: Boxplots of exports data.

Environmental Data (logs)

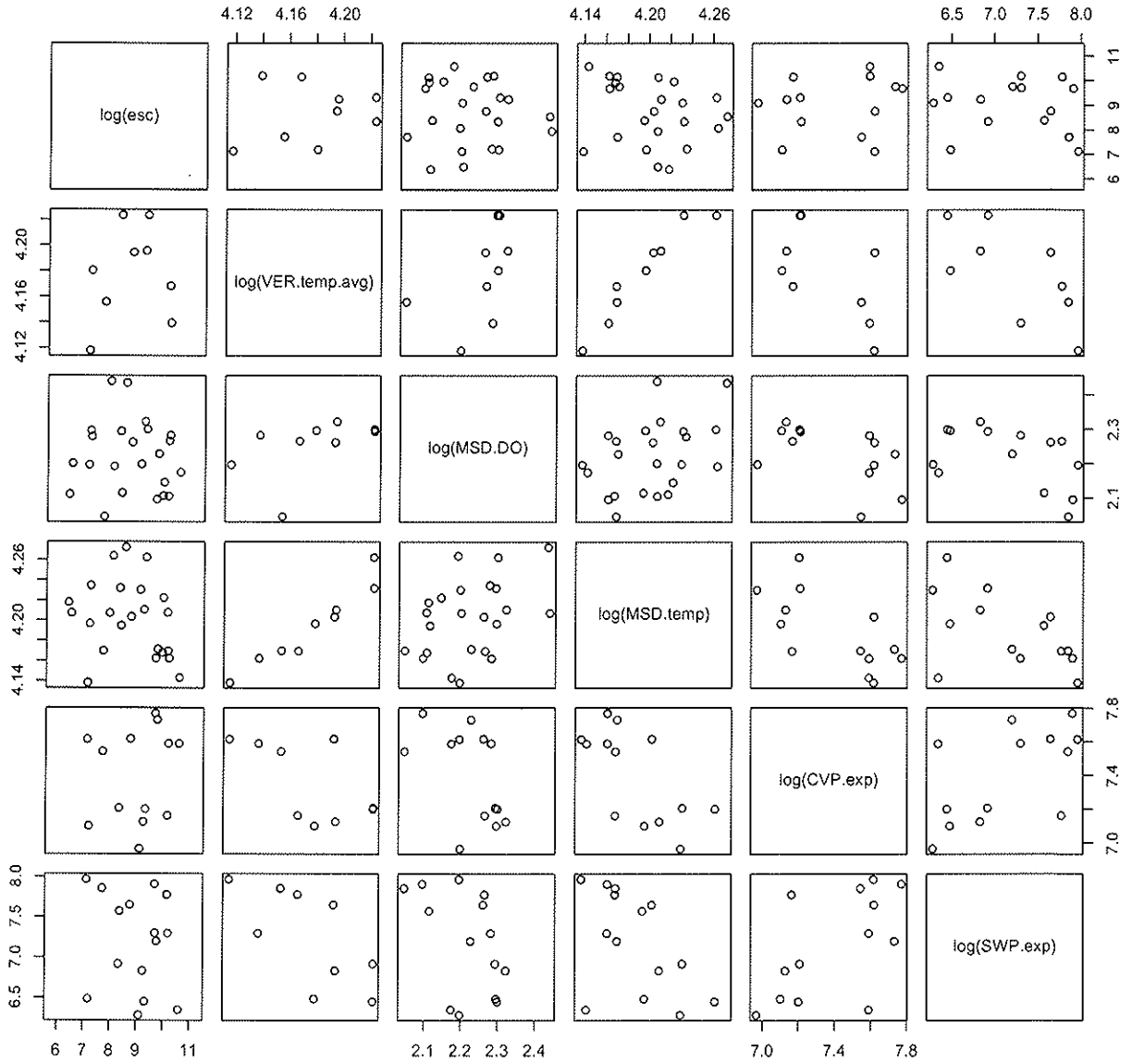


Figure 14: Scatterplots of SJR escapement and Mar. 15 - June 15 averages for other environmental data, on the log scale.

Key points regarding the DFG's Salmon Survival Model, a.k.a., San Joaquin River Salmon Population Model (Attachment 2).

1. The model's authors have neither validated its predictions using real data, nor attempted to quantify the uncertainty inherent in the modeling process. Their only evaluation of the model ignores uncertainty in its parameter estimate, assumes that the underlying form of their model is true and compares its predictions only with historical data that were used to fit the model itself, rather than relying upon more recent data.
2. In order to investigate the previous point, we have performed an analysis of the model and found its predictions to be highly unreliable because the size of its likely variability is typically larger than the prediction itself, making the prediction of no practical use.
3. Each of the component regression models in the "plug-and-play" model appears to have major violations of its distributional assumptions, making the models' behavior unpredictable.
4. The overall model is formed by "chaining together" the component regression models, feeding each model's output into another's input. This is not an accepted statistical modeling approach because it can lead to amplified errors and uncertainty in the overall model output which are difficult to quantify.
5. The model neglects available data on factors much more likely to affect salmon population than those it uses, escapement and flow. These include water temperature, ocean conditions, ocean survival, ocean predation and harvesting. Harvesting seems especially important for modeling purposes since it is subject to direct measurement and control.

Attachment 2: Report on the San Joaquin River Salmon Population Model

Gary Lorden, Ph.D.

Jay Bartroff, Ph.D.

Lordenstats

September 14, 2012

This report lists some statistical concerns with the San Joaquin River Salmon Population Model (“the model”), version 1.6 (Marston 2005; revised in Marston and Hubbard 2008). Our overall conclusion is that we are highly skeptical of any predictive value of the model and we strongly caution against its use for predicting Salmon cohort production. We are led to this conclusion by the following key points, which are expanded upon below.

1. Each of the component regression models appears to have major violations of its distributional assumptions, making their behavior unpredictable when used in this way.
2. The overall model is formed by “chaining together” the component regression models by feeding outputs of component models into other models. This is *not* an accepted statistical modeling approach because it can lead to amplified errors and uncertainty in the model output that are difficult to quantify, which the model’s authors have not addressed.
3. The model’s authors have neither validated the model’s predictions against real data, nor attempted to quantify the uncertainty inherent in their modeling process. The authors’ only evaluation of the model ignores uncertainty in its parameter estimates, assumes that the underlying form of their model is true, and compares its predictions only with historical data that is used to fit the model itself.
4. In order to investigate the previous point, we have performed an analysis of the model and found its predictions to be highly unreliable because the size of its likely variability is often larger than the prediction itself, making the prediction of no practical use.
5. The model neglects available data on factors much more likely to affect Salmon population than those it uses, escapement and flow. These include water temperature, ocean conditions, ocean survival, ocean predation, and harvesting. Harvesting seems especially important for modeling purposes since it is subject to direct measurement and control.

1 Overall Approach of Model

1.1 Description of the Model

The model has been described as an attempt to predict the number of Salmon smolts at Chipps Island in a given brood year that later return as spawners, referred to as “cohort production,” by chaining together the following three regression models.

The Mossdale Smolt Production model is a Poisson regression of Mossdale smolts counts on flow and escapement, whose expected (i.e., average) value is described by the following equation:

$$\log(EY^{(1)}) = \beta_0^{(1)} + \beta_1^{(1)} x_1^{(1)} + \beta_2^{(1)} \log(x_2^{(1)}),$$

where $Y^{(1)}$ is the Mossdale smolt count, $x_1^{(1)}$ is the average Vernalis flow over the period March 15 through June 15, and $x_2^{(1)}$ is escapement from the previous Fall.

The Delta Survival model is a logistic regression model which, by the authors’ own description, is applied to the fractional data representing the proportion of smolts that survive from Mossdale to Chipps Island, whose expected value is described by the following equation:

$$\log\left(\frac{EY^{(2)}}{1 - EY^{(2)}}\right) = \beta_0^{(2)} + \beta_1^{(2)} x_1^{(2)} + \beta_2^{(2)} x_2^{(2)} + \beta_3^{(2)} x_1^{(2)} x_2^{(2)},$$

where $Y^{(2)}$ is the proportion of smolts that survive from Mossdale to Chipps Island, $x_1^{(2)}$ is daily Vernalis flow over the period March 15 through June 15, and $x_2^{(2)}$ is either 1 or 0 according to whether the HORB was in place during that period or not.

The Cohort Production model is another logistic regression model applied to fractional data, where here the fraction is the proportion of Chipps Island smolts that become spawners, whose expected value is described by

$$\log\left(\frac{EY^{(3)}}{1 - EY^{(3)}}\right) = \beta_0^{(3)} + \beta_1^{(3)} \log(x_1^{(3)}),$$

where $Y^{(3)}$ is the proportion of smolts at Chipps Island that survive and later return as spawners, $x_1^{(3)}$ is the number of Chipps Island smolts.

The above regression coefficients $\beta_i^{(j)}$ are estimated (“fitted”) using historical data, and then the model is used for prediction of cohort production as follows. Given values of average Vernalis flow $x_1^{(1)}$, previous year’s escapement $x_2^{(1)}$, daily Vernalis flow $x_1^{(2)}$, and HORB in/out $x_2^{(2)}$, predicted values of the Mossdale smolt production $\hat{Y}^{(1)}$ and the Delta Survival fraction $\hat{Y}^{(2)}$ are generated using these values, the estimated coefficients, and the above equations. Finally, in the Cohort Production model, the predicted value $\hat{Y}^{(3)}$ of the cohort production fraction is generated using the number of Chipps Island smolts given by $x_1^{(3)} = \hat{Y}^{(1)}\hat{Y}^{(2)}$, the product of the outputs of the two previous models. Finally, the predicted value of cohort production is given by $\hat{Y}^{(1)}\hat{Y}^{(3)}$, the product of the predictions of the first and third models.

1.2 Criticism of the Model

The authors of the model claim that the predictions generated in this way seem to track the historical data on cohort production reasonably well. However, when making this comparison they do not highlight the simple fact that the model being used to make these predictions use exact numbers they are predicting as “inputs” since the historical cohort production data is used in the first step of the process described in the previous paragraph. In other words, it is not surprising that the resulting model can replicate well the historical values of estimated cohort production since those same values are used in deriving the predictions. In fact, a trivial model with the same inputs as the proposed model can replicate estimated cohort production *exactly* by simply outputting the input value for the desired year. Clearly that approach would tell us nothing meaningful. This fact should be kept in mind when assessing the proposed model, or any model of this type.

The predictions of any model of a complex system, which the San Joaquin River Salmon population certainly is, that incorporates just a few types of data must be viewed with skepticism. One of the most salient features of the current model is that it completely ignores many available sources of data which are likely to have a much stronger and more direct effect on cohort production than just escapement and flow, i.e., water temperature, ocean conditions, ocean survival, ocean predation, and harvesting. The latter is a particularly striking omission since it is a human activity that can be accurately measured and has a direct and immediate effect on fish populations.

Nonetheless, the uncertainty of any model’s predictions should be presented clearly, as well as the assumptions made about the data used to build it and “goodness of fit”-type evaluations of whether these assumptions appear to hold with the data. Unfortunately, these issues are not addressed for the current model at all. The uncertainty in the predictions of the proposed model and the model’s assumptions are not discussed by the authors and no goodness of fit tests appear to have been performed on the aggregate model, other than to look at its average output, which can easily be misleading as pointed out in the previous paragraph. The accuracy of such a model cannot be assessed without considering this uncertainty; at the very least, we suggest including confidence bands for the predictions of the model which include the uncertainty inherent in each “link” in the chain of models. In Section 3 we have performed such calculations for this model and found that the confidence intervals are so wide – even including negative numbers for a prediction of fish counts – that the predictions of the model have no utility. Perhaps the structure of the “chained together” regression models obscured the effect of the combined uncertainties of each of the individual models on the model as a whole, which occurs when the outputs of the Mossdale Smolt and Delta Survival models are used as inputs for the Cohort Production model. This technique seems to have caused a “multiplicative” effect on the predictions’ uncertainties estimated in Section 3. This violates a fundamental assumption of the regression models of the type used, that the inputs (i.e., independent variables) are fixed quantities that are known exactly (e.g., can be measured with minimal measurement error). When this does not hold, a different type of regression model known as “errors in variables” models should be used. In the current situation, the outputs of the Mossdale Smolt and Delta Survival models have considerable uncertainty under even the most generous assumptions, so it seems clear that these assumptions are violated and a different statistical approach is required.

Some other problems with the “chained together” structure of the model is that it creates dependencies between the errors in the models’ outputs and inputs, which is another violation of assumptions of the component regression models which can cause their behavior to be unpredictable because of increased variability in their output and amplified error propagation. Further, the component regression models themselves each suffer from “poor fits,” evident through the presence

of outliers to which the types of models used are well-known to be non-robust, small coefficient of determination R^2 , a few overly influential observations, poor quantile-quantile plots, and poor goodness of fit scores. In the next section we give more details on these and other problems with the component regression models.

2 Criticism of Individual Components of the Model

2.1 Mossdale Smolt Production Model

There are three main areas of concern regarding the Mossdale Smolt Production model which likely affect its output and the output of the combined model.

- 1. Weak relationship between flow and smolt production:** The scatterplots in Figure 1, the data used to fit the Mossdale Smolt Production model, show a weak relationship between flow and smolt production, which violates a fundamental assumption of the model.
- 2. Overly influential observations:** The data shows a small number of overly influential observations with high flow levels inflating “upward” trend; see, for example, the three outlying values in the flow vs. smolts scatterplot in Figure 1. Outliers such as these are well-known to inflate the estimated relationship between the response variable (smolts) and the input (flow) in generalized linear models like the Poisson regression model used here. Evidence of this is that the model fit changes dramatically when these are removed: See Figure 2, which shows a linear model fit to the full data set (in black) and the quite different fit to the data set with the outliers removed (in red). These two fits would produce vastly different outputs for the “chained together” model.
- 3. Overdispersion:** This model fit also appears to suffer from *overdispersion*, i.e., when the variance of data is larger than the average value of the data. A fundamental assumption of the Poisson regression model is that the mean response is exactly equal to the variance of the response, which appears to be violated by the data; see the last row or column of Figure 1.

2.2 Delta Survival Model

An area of concern regarding the Delta Survival model which likely affects its output and the output of the combined model is that, by the authors’ description, an incorrect use of a logistic regression model: the model was fit with proportions (i.e., percent survival estimates from coded wire tag studies) rather than binary data (i.e., data comprised of two outcomes, either a 0 or 1). The two methods are only equivalent if the total number of smolts leaving Mossdale was exactly the same for each data point (i.e., CWT releases groups were all the same size), which is not true for the data used (i.e., CWT release groups were variable within and between years). This discrepancy can be corrected by a re-weighting of the data, but this did not occur in the authors’ description. Using the Delta Survival model with fractional data violates the fundamental distributional assumption of logistic regression models.

Data in Mossdale Smolt Production Model

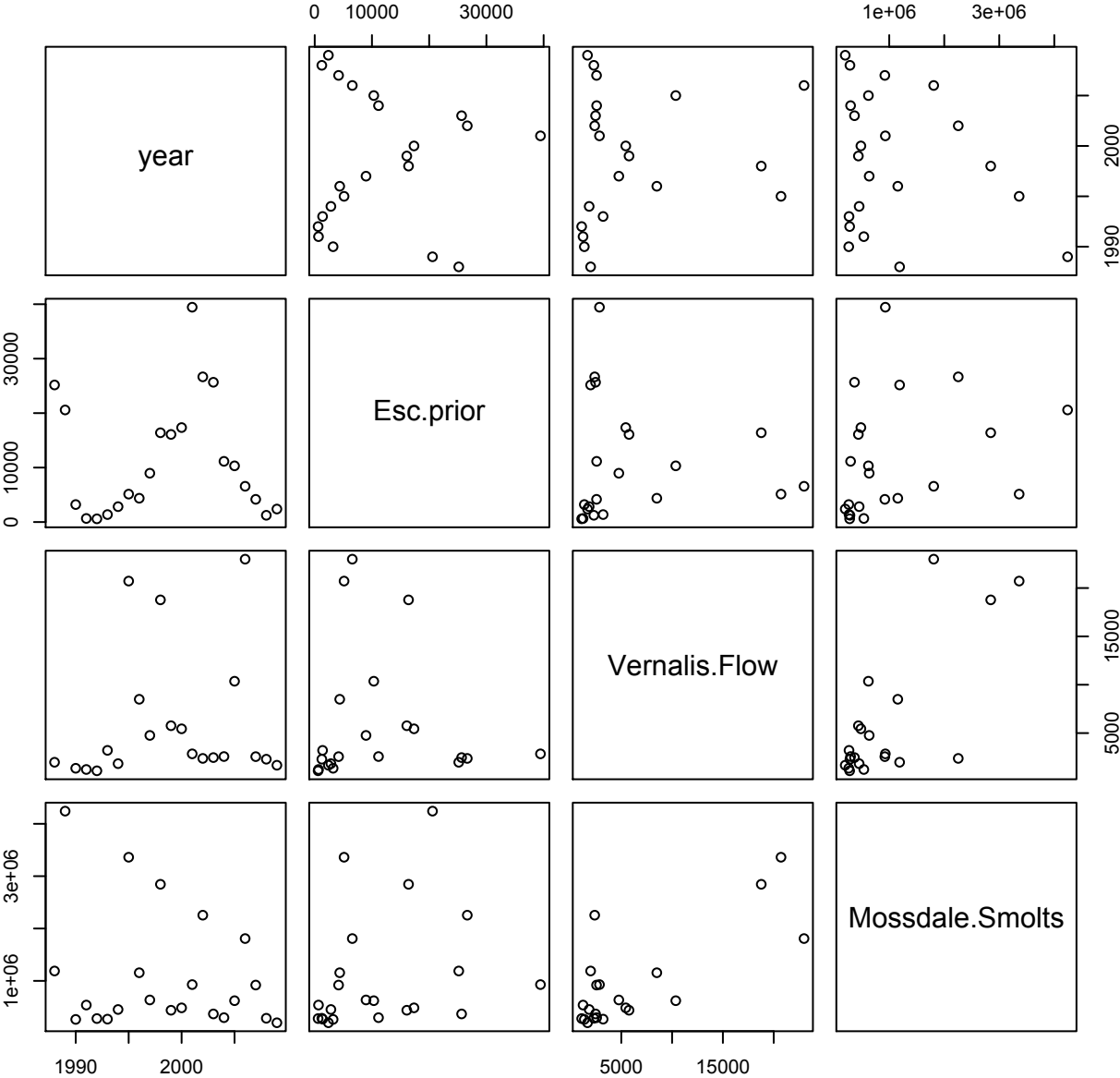


Figure 1: Data used to fit the Mossdale Smolt Production Model

Regression Model for MD Smolt

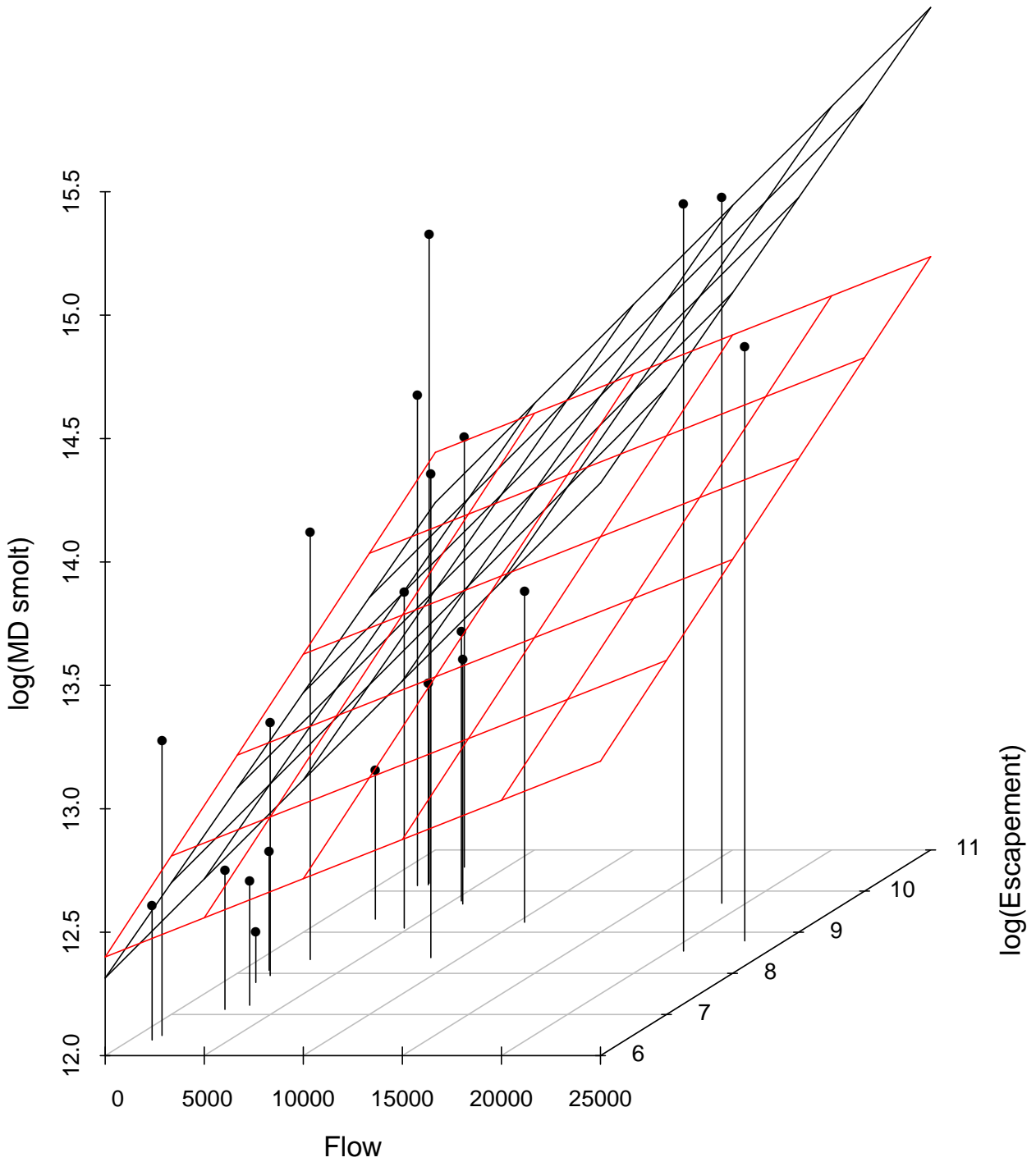


Figure 2: Linear models fit on Mossdale Smolt Production Model data: Model fit on full data is in black, and model fit on data with outliers removed is in red.

2.3 Cohort Production Model

There are two main areas of concern regarding the Mossdale Smolt Production model which likely affect its output and the output of the combined model.

- 1. Fundamental distributional assumption violated:** Like the Delta Survival model, the Cohort Production model appears to also have been fit with fractional data of different sizes, as described in the previous paragraph, which violates the distributional assumption of the logistic regression model.
- 2. Problematic use of $N_{Chippis}$:** The “independent” variable (i.e., the input) in this model is an estimate $N_{Chippis}$ of the number of Chippis smolts ($N_{Chippis} = \hat{Y}^{(1)}\hat{Y}^{(2)}$ in the notation of Section 1.1), whereas the output of the overall model itself gets multiplied by $N_{Chippis}$ to give the final prediction of the chained model. This is equivalent to a simple transformation of $N_{Chippis}$, and it would more accurately be described in that way. Moreover, the “dependent” variable in this model is the fraction of the number $N_{Chippis}$ of smolts at Chippis Island that return as spawners, so $N_{Chippis}$ is also the denominator in the response variable data used to fit the model. $N_{Chippis}$ itself is the output of the Mossdale Smolt Production model, and hence has an inherent uncertainty beyond what would already be present in a measured data point. This likely high level of uncertainty thus contributes uncertainty to the output of the Cohort Production model in three different ways.

3 Our Analysis of the Model

We have performed an analysis of the model in order to address two fundamental questions:

1. How accurate are the model’s point predictions of Salmon cohort production when they are not calculated from the data it is asked to predict?
2. How much uncertainty is there in the model’s predictions of Salmon cohort production? In other words, if the same modeling process were repeated with fresh data of the same type, how much would the predictions change?

To answer these questions we have repeated the authors’ modeling process but using only the data available prior to a given year n , and then used the resulting model to predict cohort production for year n and compare it with the measured value of cohort production, which we call the “true value.” We have done this for years $n = 1996$ through 2001, which is the widest range of years possible using the data provided by the authors. This validation process is used because it mimics how the model would be used in practice: The model would be fit using all available data up to the present, and then it would be used to predict future cohort production, presumably under various scenarios.

In addition, we have used the standard statistical method known as the Jackknife to assess the uncertainty in the model’s outputs. This method estimates uncertainty by repeating the parameter fitting process while leaving out one data point at a time, and seeing how much the predictions of the resulting model change. This is repeated for all possible data years and the information is combined to give an overall estimate of the likely variability in the model’s predictions were the modeling repeated with fresh data of the same type. This is known as the *standard error*, and is the accepted way of assessing variability of a prediction. For example, for normally distributed data,

the average of the data points will fall within one standard error of the population’s true mean roughly 2/3 of the time, within two standard errors roughly 95% of the time, and within three standard errors more than 99% of the time.

The results of this analysis are given in Table 1, wherein the predictions of cohort production differ from the true values by large amounts. For example, for 1997 the predicted cohort production of 8,713 is less than half the true value of 18,221, while in 2001 the predicted cohort production of 80,524 is nearly 6 times the true value of 13,763. Inaccuracy of these point predictions aside, the standard errors computed via the Jackknife method are the same order of magnitude, and larger in most cases, than the predictions themselves. This makes the predictions of the model of essentially no value since it means that if this method were repeated with similar data (e.g., *future* data), then the model’s new predictions could likely be very small, or twice as large.

Table 1: True and predicted values of cohort production, and the predictions’ standard errors, using a model fit only with data available prior to the given year.

Year	Cohort Production		
	True Value	Model’s Prediction	Standard Error
1996*	7,164	6,789	10,953.5
1997	18,221	8,713	14,309.4
1998*	48,491	40,859	70,568.5
1999	18,471	18,570	13,791.5
2000	21,608	55,234	75,696.2
2001*	13,763	80,524	77,941.9

In addition to the overwhelmingly large standard errors in Table 1, the uncertainty in the model’s predictions can further be seen by examining more closely the Jackknife method used to compute these standard errors which, as mentioned above, “re-fits” the model by leaving out one data point at a time. For example, the data made available by the authors includes (estimates of) the number of smolts at Chipps Island and the resulting total cohort production for the years 1988 through 2001. Focusing on the model’s predicted cohort production of 80,524 for the year 2001, in Table 2 we give what the value of this prediction would be if a single year’s data were left out of the Cohort Production model, with the Mossdale Smolt and Delta Survival models fit using their full data sets as usual. For example, if the 1988 data is left out when fitting the model, the predicted cohort production jumps from 80,524 to 100,739; if 1989 is left out instead then this number is still higher at 114,674; and if the following year 1990 is left out instead then the prediction falls all the way down to 45,416. Such wide swings in the predictions of the model, resulting from only the small change in the input data of leaving out one data point, are evidence of the non-robustness and instability of the modeling method and should be a caution to anyone considering its predictions.

This same technique can be used to analyze the model’s variability from a slightly different angle, to answer the question, “How different would the model’s sequence of predictions forward in time be if a single year’s data was left out in the model-fitting process?” A robust modeling technique should not be overly affected by a small change in data like this, but Figure 3 shows that

*For these years, multiple flow measurements for the Delta Survival model were available in the data, making multiple model predictions possible. In each case, the flow measurements and model predictions were very similar to the others of the same year, so we have included only one in the table for each year.

Table 2: Model's predictions of cohort production for 2001 when one year is left out of fitting the Cohort Production component model. The original prediction with no year left out is 80,524.

Year Left Out	Model Prediction
1988	100,739
1989	114,674
1990	45,416
1991	76,414
1992	97,236
1993	80,465
1994	81,097
1995	95,786
1996	102,705
1997	64,466
1998	82,678
1999	63,099
2000	56,698

this model's predictions change drastically when a single year is omitted in the coefficient-fitting steps. For example, for predicting the cohort production in 1997, the prediction grows by 113% if the 1988 input data are omitted, while the prediction falls 74% if the 1990 input data are omitted, a 187% swing.

% Change in Prediction When Leaving Out One Year in 1988–1992

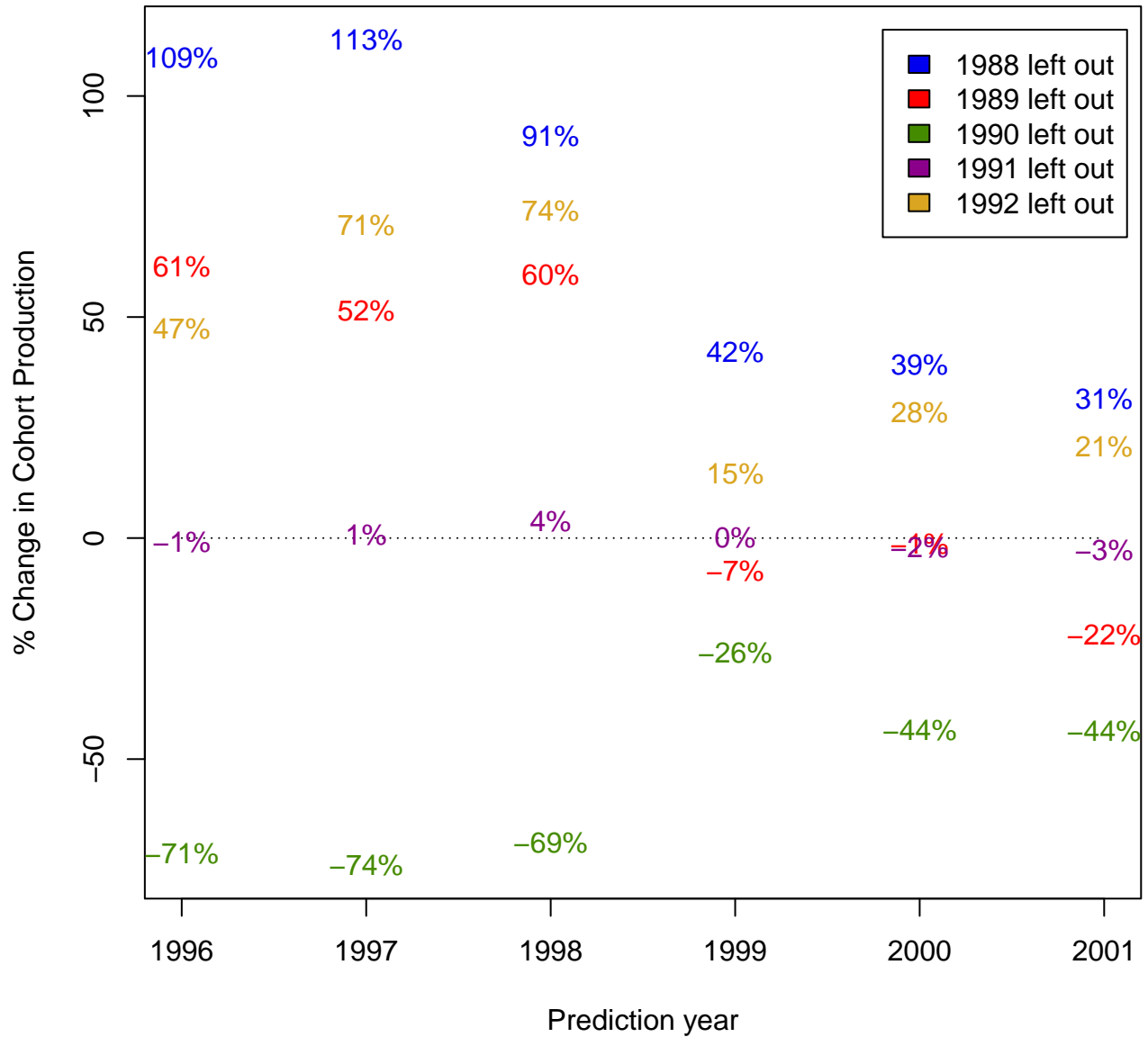
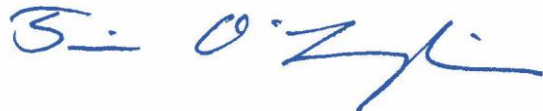


Figure 3: The percent change in the model’s cohort production when a single year is left out of the data fitting the regression coefficients.

Charlie Hoppin, Chairman
State Water Resources Control Board
October 26, 2012
Page 5

Very truly yours,
O'LAUGHLIN & PARIS LLP



TIM O'LAUGHLIN

TO/tb
cc: San Joaquin Tributaries Authority

References

*Demko, D., M. Palmer, S. Snider, A. Fuller, S. Ainsley, M. Allen, and T. Payne. 2010. Comments pertaining to the “Scientific Basis for Developing Alternate San Joaquin River Delta Inflow Objectives” described in the State Water Resources Control Board’s October 29, 2010, *Draft Technical Report on the Scientific Basis for Alternative San Joaquin River Flow and Southern Delta Salinity Objectives*. Submitted to the State Water Resources Control Board on behalf fan Joaquin River Group Authority, December 6, 2010.

http://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/bay_delta_plan/water_quality_control_planning/comments120610/michele_palmer.pdf

*Lorden, G. and J. Bartroff. 2010. Report on flow vs. escapement model and environmental data: Lordenstats, December 1, 2010. Report provided in Appendix 1 of Demko et al. 2010.

http://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/bay_delta_plan/water_quality_control_planning/comments120610/michele_palmer_attachment1.pdf

Lorden, G. and J. Bartroff. 2012a. Report on flow vs. escapement model and environmental data. April 16, 2012. Report provided in Attachment 1 of this memorandum.

Lorden, G. and J. Bartroff. 2012b. Report on the San Joaquin River Salmon Population Model. September 14, 2012. Report provided in Attachment 2 of this memorandum.

Marston, Dean. 2005. *FINAL DRAFT*: San Joaquin River Fall-run Chinook Salmon Population Model. Submitted to State Water Resources Control Board for Periodic Review of the Bay-Delta Water Quality Control Plan.

Marston, D. and A. Hubbard. San Joaquin River salmon population model. SWRCB SJR Flow Workshop Sept. 17, 2008.

http://www.waterrights.ca.gov/baydelta/docs/sanjoaquinriverflow/dfgpresentation_salmon.pdf

SWRCB [State Water Resources Control Board]. 2012. Technical report on the scientific basis for alternative San Joaquin River flow and Southern Delta salinity objectives. February 12, 2012. 202 pp.

* **Written comments previously submitted on December 6, 2010, to the SWRCB regarding SWRCB’s Draft Technical Report (links provided in citations).**